Discovery and Scoring of Protein Interaction Subnetworks Discriminative of Late Stage Human Colon Cancer*Ⓢ

Rod K. Nibbe‡§¶, Sanford Markowitz‖, Lois Myeroff‖, Rob Ewing§, and Mark R. Chance§**

We used a systems biology approach to identify and score protein interaction subnetworks whose activity patterns are discriminative of late stage human colorectal cancer (CRC) *versus* control in colonic tissue. We conducted two gel-based proteomics experiments to identify significantly changing proteins between normal and late stage tumor tissues obtained from an adequately sized cohort of human patients. A total of 67 proteins identified by these experiments was used to seed a search for protein-protein interaction subnetworks. A scoring scheme based on mutual information, calculated using gene expression data as a proxy for subnetwork activity, was developed to score the targets in the subnetworks. Based on this scoring, the subnetwork was pruned to identify the specific protein combinations that were significantly discriminative of late stage cancer *versus* control. These combinations could not be discovered using only proteomics data or by merely clustering the gene expression data. We then analyzed the resultant pruned subnetwork for biological relevance to human CRC. A number of the proteins in these smaller subnetworks have been associated with the progression (*CSNK2A2*, *PLK1*, and *IGFBP3*) or metastatic potential (*PDGFRB*) of CRC. Others have been recently identified as potential markers of CRC (*IFITM1*), and the role of others is largely unknown in this disease (*CCT3*, *CCT5*, *CCT7*, and *GNA12*). The functional interactions represented by these signatures provide new experimental hypotheses that merit follow-on validation for biological significance in this disease. Overall the method outlines a quantitative approach for integrating proteomics data, gene expression data, and the wealth of accumulated legacy experimental data to discover significant protein subnetworks specific to disease. *Molecular & Cellular Proteomics 8:827–845, 2009.*

A fundamental presumption of the *-omics* revolution is that high dimensional data sets resulting, for example, from proteomics and genomics experiments should be integrated with functional annotations to give a more complete account of the cellular changes underlying the etiology of human disease. Nevertheless the accumulation of specific gene annotations and experimental protein or gene expression data is presently outpacing data integration. Network modeling of protein-protein interactions provides a context for such data integration (1–3). These modeling approaches can build networks using databases created from literature curation, inference by homology, high throughput data, or a combination of these (4). Network generation, analysis, and modeling are clearly fundamental to a new generation of systems biology approaches that promise an improved understanding of the causes of human disease as well as providing novel biomarkers of its progression.

We undertook a systems biology approach to identify protein "signatures" that were significantly discriminative of late stage human colorectal cancer (CRC)[1] *versus* control. CRC continues to be the second leading cause of cancer death in adult Americans (5). Although a great deal of research is focused on the *early* detection of CRC, comparably less attention has been paid to understanding the pathophysiology of a late stage (Duke's D) phenotype. As the prognosis of a late stage diagnosis is significantly poorer ($<$10% long term survivability (5)) than one following early detection ($>$90% (5)), identifying significant network-level changes in a late stage cohort holds the possibility of more clearly elucidating the mechanisms of tumorigenesis specific to this phenotype.

Proteomics studies of CRC using tumor and adjacent normal tissue obtained by biopsy from human patients have been conducted (6–8). Even more studies have profiled protein expression changes in colon cancer cell lines (9–13). However, these tissue-based studies have used either a sample cohort of mixed pathologic stage or a cohort size that was smaller than optimal. As colon cancer is a disease that

From the Departments of ‡Pharmacology and ‖Medicine, §Case Center for Proteomics and Bioinformatics, and **Howard Hughes Medical Institute, Case Western Reserve University, Cleveland, Ohio 44106

Received, September 9, 2008, and in revised form, November 12, 2008
Published, MCP Papers in Press, December 19, 2008, DOI 10.1074/mcp.M800428-MCP200

[1] The abbreviations used are: CRC, colorectal cancer; MI, mutual information; 2D, two-dimensional; TEMED, *N*,*N*,*N*′,*N*′-tetramethylethylenediamine; ABC, ammonium bicarbonate; H, hypothesis; CDF, cumulative distribution function; TRAIL, tumor necrosis factor-related apoptosis-inducing ligand; TCF, T-cell factor; PDGFR, platelet-derived growth factor receptor; PLK, Polo-like kinase; CCT, chaperone containing t-complex protein; LTQ, linear Trap quadrapole.

progresses over a number of years and is marked by distinct pathologic stages of increasing severity (Duke's A–D), it is reasonable to expect changes in the proteome that are associated with particular stages of disease. Hence a cohort of homogenous pathology may improve the detection of stage-specific changes. Further we expected that the most dramatic changes in protein expression, in terms of quantity and magnitude, would be detectable between control and tumor using a statistically robust late stage cohort.

Fig. 1 outlines our overall experimental design and emphasizes an integrated -omics approach to understanding the pathophysiology of stage D colon cancer. We began by quantitative proteomics profiling of a late stage CRC cohort and used the differentially expressed proteins to seed a search for protein interaction subnetworks possibly involved in this disease (Fig. 1, *steps 1–11*). Our list of differentially expressed proteins was imported into a bioinformatics data mining tool that permits a search of a database comprised of tens of thousands of manually curated protein-protein interactions (14). This provided us with a list of subnetworks that we were able to rank by significance and reduce to a subsequently manageable number on which to focus our analysis (Fig. 1, *steps 12–14*).

We seeded our initial search for subnetworks with proteomics data (*versus* transcriptomics data) because changes in both protein expression and isoform abundance provide the most direct functional readout of the cell. As such, we expected our seed proteins would represent "fence posts" within the subnetworks, each with one or more functional roles. These subnetworks represent expansions of the cancer proteome in the sense that the algorithm we used (see "Experimental Procedures") builds an extended interactome comprised of many targets around a smaller set of seed proteins. This extended interactome, although it provides clues to the regulatory connections that drive the observed abundance changes, is merely qualitative and inferential. Thus, a potential criticism of this type of approach is that a small set of proteins that is potentially causative with respect to a disease state has simply been expanded to a larger set whose members may or may not be important to the phenotype. If the myriad of protein interaction networks (and there are many tools to build interactomes from a set of seed targets) are to be useful to researchers for informing new hypotheses, new methods are needed to quantitatively evaluate the significance of the targets within the subnetwork that is generated. To address this, we adapted a method described by Ideker and co-workers (15) to score our subnetworks, and then we systematically searched within each one using the metric of mutual information (MI) to identify statistically significant combinations of proteins that were highly discriminative of stage D cancer *versus* control. Gene expression data were an ideal basis for scoring our subnetworks because of their complete coverage, *i.e.* we were able to assign an mRNA expression value to every target.

Overall our guiding principle is that protein is the immediate effector of phenotype. Therefore, profiling changes in the proteome is likely to provide the most direct evidence for cellular changes causing, or resulting from, a disease. However, proteomics experiments typically have incomplete coverage of the proteome. In particular, gel-based expression experiments are most likely to detect high abundance proteins. These high abundance targets may include subnetwork nodes that participate in larger subnetworks of protein interactions and may also be regulated at the level of transcription. If so, patterns of mRNA expression can be useful for discriminating between disease and non-diseased states within these "discovered" subnetworks. Of course, mRNA expression data have the characteristic of whole genome coverage, and these data can enable queries for subnetworks of interest that are "saturated." Here we present an integrated approach to cancer biology, one that shows how proteomics data, genomics data, and a vast database of legacy experimental data can be integrated with MI scoring schemes to reveal protein signatures significantly discriminative of disease. The signatures are useful for focusing follow-on experiments to verify their functional role in a disease phenotype. In addition, our approach is very general for use with existing public data sets as well as newly generated data and can be applied in the context of multiple types of protein interaction networks.

EXPERIMENTAL PROCEDURES

Sample Preparation

pI 3–10 Experiment—Tissue samples were procured from a human tissue repository at the Case Comprehensive Cancer Center (supplemental Data S2). In addition to a tumor biopsy during surgical resection, a normal biopsy adjacent to the patient's tumor was also taken, typically >10 cm from tumor. Validation of the tissue as normal or tumor (including stage of the tumor) was performed by a pathologist. Tissues were immediately frozen and stored at −80 °C. A 50-mg sample provided an adequate mass of protein for the 2D DIGE experiment. The tissue was weighed and placed in lysis buffer (4% CHAPS, 7 M urea, 2 M thiourea, 30 mM Tris) on ice, and the cells were disrupted by a three-cycle sonication protocol in a 4 °C cold room. A protease inhibitor mixture (Sigma-Aldrich, catalog number P8340) and a wide spectrum phosphatase inhibitor (Roche Applied Science) were added to the buffer at the manufacturer's suggested concentration to inhibit protein degradation and dephosphorylation, respectively. The homogenate was centrifuged at 12,000 rpm for 10 min, and the protein fraction was withdrawn by pipette. Protein concentration was quantified by a colorimetric assay, similar to the Bradford assay, using the 2D-Quant kit (GE Healthcare). Aliquots were stored at −80 °C.

pI 4–7 Experiment—Protein fractions from the prior experiment were thawed and cleaned with the 2D-Cleanup kit (GE Healthcare), and the concentration was redetermined as before. Aliquots were re-stored at −80 °C.

2D Gel Electrophoresis

We used the 2D DIGE system available from GE Healthcare (formerly Amersham Biosciences) described by Marouga *et al.* (39). This system provides two distinct advantages over conventional 2D PAGE. First, it allows for up to three distinct samples to be labeled by spectrally resolvable fluorophores (CyDyes Cy2, Cy3, and Cy5) and multiplexed in a single gel. Second, by using one of these CyDyes (typically Cy2) to label a pooled sample, constituted by a proportional amount of *every* sample in the experiment, the Cy2 dimension is useful as an internal standard. This internal standard is crucial in the image analysis phase to a confident assessment of real biological variation from gel to gel as distinct from changes arising from variance in protein loading. For the purpose of detection by image analysis, 50 μg of protein is sufficient for labeling by each of the CyDyes. Additionally gels intended to be used for spot excision were loaded with an additional 350 μg of an unlabeled, pooled sample sufficient for tryptic digestion and detection of the peptides by LC-MS^2.

First Dimension—Each *minimal* CyDye was reconstituted in fresh *N*,*N*-dimethylformamide, and a 400-pmol quantity was used to label 50 μg of protein at pH 8–9. Cy2 was used to label the pooled internal standard as described above. Cy3 and Cy5 were used to label the normal and tumor samples, and we alternately swapped the dyes on subsequent sample pairs to alleviate dye-specific effects that could bias image analysis. Labeling proceeded for 30 min in the dark and was quenched with 10 mM lysine. Samples were then mixed with an equal volume of 2× sample buffer (8 mM urea, 4% CHAPS, 2% DTT, 2% Pharmylyte, pH 3–10 or 4–7, non-linear), placed on ice for 10 min, then loaded onto non-linear pH 3–10 (or 4–7) Immobiline DryStrips (GE Healthcare), placed in a strip holder, and focused with an IPGphor system using a step gradient protocol ranging from 30 to 8000 V for approximately 27 h. The strips were then stored at −80 °C, ready for the second dimension. Additionally for the first experiment (pI 3–10), two pooled, unlabeled 350-μg samples were prepared and focused separately to be subsequently separated in the second dimension on *separate* gels intended for spot excision. By contrast, for the second experiment (pI 4–7), the unlabeled, pooled sample was mixed with the labeled samples and run on the *same* gel. This is possible because the Deep Purple gel stain (GE Healthcare) we used to stain the unlabeled sample is spectrally resolvable from the CyDye fluorophores. This reduces the number of gels required for the experiment.

Second Dimension—For separation by molecular weight we used the Ettan DALT Twelve apparatus. The DryStrips were rehydrated in 10 μl of re-equilibration buffer (8 M urea, 100 mM Tris-HCl, pH 6.8, 30% glycerol, 1% SDS, 45 mg/ml iodoacetamide (to reduce streaking)) for 10 min, laid across the top of a homogeneous 12.5% polyacrylamide gel "sandwiched" between two glass plates submerged in running buffer (40% bisacrylamide, 1.5 Tris, 10% SDS, 10% ammonium persulfate, 10% TEMED), and then covered with a 0.5% agarose solution. Separation proceeded at 15 °C at 0.5 watts/gel and then 1.0 watt/gel for 15 h. Separation was stopped when the bromphenol dye front reached the bottom of the gel.

Gel Fixation—Gels to be used for spot excision were previously secured to one glass plate in the "sandwich" with silane (Bind-Silane, GE Healthcare). After the experiment was stopped, these gels were fixed in 50% methanol and 7.5% acetic acid and subsequently stained with Deep Purple Total Protein Stain (GE Healthcare).

Image Analysis

Gels were scanned using a Typhoon 9400 variable mode imager (GE Healthcare). During this phase each CyDye fluorophore is independently excited by laser light specific to its particular excitation spectrum. Emission sensitivity, *i.e.* photo multiplier tube, was adjusted until the most intense spot on the gel approached saturation. This tuning was performed at a 1000-μm resolution, and once the photo multiplier tube value was optimal, a final high resolution scan was performed at 100 μm. The gel(s) intended to be used for spot excision was poststained with Deep Purple (GE Healthcare), imaged using 532/560 nm wavelength light, and then stored in the dark at 4 °C. In general, by following the recommended settings for the Typhoon imager, our experience indicates that dye-specific biases in spot intensity are eliminated or reduced below significance. After imaging the three fluorophores for each gel, the images were imported to the DeCyder image analysis software (GE Healthcare) for spot detection, spot matching (intragel), and determination of statistically significant biological variation (intergel) based on the measurement of relative abundance change after background subtraction and normalization to the internal standard. Typically about 90% of the spots on a gel will fall within 2 standard deviations of the mean and not show a significant -fold change (the null hypothesis), although this is highly sample-dependent.

Image analysis is a time-consuming part of this experiment. A statistically significant spot is one whose mean -fold change is greater than or equal to ±50% (depending on statistical power) and paired *t* test is less than or equal to 0.05. Each spot that passes significance must then be manually checked to ensure that it is likely a protein spot and not a gel artifact. Satisfying these criteria, a pick list is generated and exported to the software controlling the Ettan robotic spot picker (GE Healthcare). Spots were excised with a 3-mm core from the poststained gel and loaded to a 96-well plate for digestion.

In-gel Digestion

Excised gel plugs were washed four times for 10 min with 50 μl of both 25 mM ammonium bicarbonate (ABC) and 50% ACN, removing the liquid between each wash. 10 mM DTT freshly prepared in 30 μl of 25 mM ABC was added to each gel plug. The samples were then incubated for 45 min at 56 °C. Following incubation, gel plugs were cooled at room temperature for 20 min. The DTT was removed, and 30 μl of 55 mM iodoacetamide was added. The samples were incubated in the dark for 45 min at room temperature. The iodoacetamide was removed, and samples were washed four times with 50 μl of both 25 mM ABC and 50% ACN. Gel plugs were covered with 10 μl of a 100-ng trypsin solution, incubated at room temperature for 10 min to allow absorption of trypsin, then covered with 15 μl of 25 mM ABC, and placed in a 37 °C water bath overnight for digestion. The reaction was quenched the following day with 7 μl of 1% (final concentration) formic acid. Extraction of the peptides from the gel plugs was completed by adding 30 μl of 50% ACN, 5% formic acid, vortexing for 30 min, spinning the samples, and finally sonication for 5 min.

Mass Spectrometry and Database Software

Most samples were analyzed by tandem mass spectrometry using an LTQ mass spectrometer (Thermo Electron Corp., Bremen, Germany) equipped with an Ettan multidimensional LC system (GE Healthcare). Six samples were run on a Finnigan LTQ FT hybrid mass spectrometer (Thermo Electron Corp.) operated in positive ion mode. 2.5 μl of tryptic peptides were desalted on a C_{18} pre column (PepMap 100, 300 × 5-μm particle size, 100 Å, Dionex) and then separated on a reverse-phase column (C_{18}, 75 μm × 150 nm, 3 μm, Dionex) using mobile phases A (0.1% formic acid) and B (84% acetonitrile, 0.1% formic acid) with a linear gradient of 2%/min, beginning with 100% A. Peptides were subsequently infused at a flow rate of 300 nl/min via a Pico Tip emitter (New Objective, Inc., Woburn, MA) at a voltage of 1.8 kV. Mass spectra were recorded in the ion trap, and MS^2 spectra were acquired for the five most intense ions in the LTQ using a collision energy of 35 eV and an isolation width of 2.5 Da.

Bioworks version 3.2 (Thermo Electron Corp.) using the SEQUEST software was used to search against an indexed human database with a peptide mass tolerance of 2.5 Da and a fragment tolerance of 1.0 Da. Search parameters included partial methionine oxidation, complete carbamidomethylation of cysteine, and two missed cleavage sites. Statistically significant peptides were those satisfying $p <$ 0.001 and cross-correlation (Xcorr) values of 1.9, 2.5, and 3.0 for 1+, 2+, and 3+ charged ions, respectively. The protein probability cutoff was $p < 0.001$, and each "hit" necessarily required at least three peptides for consideration with rare exceptions for low molecular weight proteins. Surviving that filter, each protein call was manually "rationalized" to the gel; that is the theoretical pI and molecular weight were compared with the observed values on the gel image. Cleavage products and post-translational modifications were considered in this step.

Statistical Power Analysis

The power of a statistical test $(1 - \beta)$ is a measure of the probability of correctly rejecting the null hypothesis, H_0, if it is false. Low power studies consequently have a higher rate of false negatives. Formally β is functionally related to sample size (n), the standard deviation of the distribution (σ), the difference in the means being tested ($\mu - \mu_0$), and the area under the standard curve at a given significance level (α) (Z_{100}).

$$n = \sigma^2 \frac{|Z_{100(1-\alpha)} - Z_{100(1-\beta)}|}{(\mu_0 - \mu_1)^2} \qquad \text{(Eq. 1)}$$

Prior to our second experiment we estimated the average spot variance (σ) by considering all spots on all 12 gels under the assumption that the source of variance was primarily biological and that experimental variance was relatively minimal. Because the samples had been prepared, labeled, and separated at once under near identical conditions we thought this assumption reasonable. Next, at a fixed level of significance ($\alpha = 0.05$) we calculated the relationship between power and fold change ($\mu_0 - \mu_1$) at three different sample levels ($n =$ 3, 6, and 12). This provided an estimate of the minimum number of paired samples required to measure a particular minimum fold change with a power of 0.8 (see supplemental Data S1).

Gene Expression Data

mRNA expression was measured by cDNA microarray on 171 human colon tissue samples of various stages of colon cancer (normal = 16, stage B = 41, stage C = 25, stage D = 50, metastatic = 39) using the Affymetrix Human EXON 1.0 ST chip. Expression values were generated with the Expression Console program from Affymetrix (Affymetrix, Santa Clara, CA) using the probe logarithmic intensity error (PLIER) algorithm to minimize the effect of outliers. Expression values for all 171 samples for select genes in our networks, plus the decoy database of 1000 genes, were generously provided to us by the Case Comprehensive Cancer Center. The decoy genes were randomly chosen from >17,800 probe sets with core evidence. The distribution of the decoy was evaluated to ensure representation across all 23 (1–22 plus X) chromosomes (supplemental Data S3) and was verified not to overlap any genes in our four networks. The data set is available upon request.

Protein Interaction Network Database and Subnetwork Build Algorithm

We used MetaCore from GeneGo Inc. (version 4.6 build 12332) to search for protein-protein interaction networks. MetaCore uses a protein interaction database comprising tens of thousands of protein interactions that have been manually curated based on a thorough reading of evidence reported in the literature. MetaCore covers 2400 journals and does not use natural language processing algorithms. In essence, the database represents a vast wealth of legacy experimental evidence that can be quickly mined in a number of ways for proteins and interactions relevant to a particular disease. These data can be usefully represented by directed graphs ("networks") that illustrate not only which proteins interact with each other but the functional nature of the interaction between them (binding, cleavage, phosphorylation, etc.). We will use the term "network" to refer to the entire database of protein interactions, and "subnetwork" will mean any network smaller than the whole network. There are a number of algorithms available in MetaCore with which to build subnetworks around a set of differentially expressed targets ("seed"). We chose an algorithm that would extend a subnetwork around our seed while minimizing the number of outgoing and/or incoming connections needed to enclose the seed in a "cloud" of interactions topologically constrained by the shortest path. As this subnetwork is likely to be very large, it is subsequently divided into smaller subnetworks by maximizing the saturation of the seed targets in each while obeying our input constraint of subnetwork size ($n = 50$). We further constrained the search by species (human) and tissue type (colon); all other prefilter options retained their default values. The end result was a list of subnetworks (13) ranked by p value and zScore. The reported p value is calculated assuming a hypergeometric distribution, and it represents the probability of a particular mapping arising by chance given the numbers of genes in the set of total networkable genes (*i.e.* genes or network objects that have at least one annotated functional interaction), all genes on maps/subnetworks/processes, genes on a particular map/subnetwork/process, and genes in our experiment. The zScore is a statistical measure of the concentration of the seed targets in the subnetwork.

$$\text{zScore} = \frac{r - n\frac{R}{N}}{\sqrt{n\left(\frac{R}{N}\right)\left(1 - \frac{R}{N}\right)\left(1 - \frac{n-1}{N-1}\right)}} \qquad \text{(Eq. 2)}$$

where N equals the total number of nodes in MetaCore database, R equals the number of the objects of the subnetwork corresponding to the genes in the import list, n equals the total number of nodes in each subnetwork generated from the import list, and r equals the number of nodes with data in each subnetwork generated from the import list. A white paper providing additional details of network construction algorithms is available upon request.

Network Scoring and Significance Tests

A flow chart of the scoring scheme is outlined in Fig. 5. We obtained global gene expression data as measured by cDNA microarray (for detail see "Gene Expression Data") for every gene product in each of the four subnetworks chosen for analysis. Importantly although the search criteria constrained each subnetwork to 50 proteins or less, some proteins in the subnetworks were complexes involving multiple gene products, subunits, or isoforms. We chose to include all these gene products when we exported the subnetwork list of genes from MetaCore. Consequently and depending on the particular subnetwork, the number of genes may exceed 50. Additionally the microarray data set also included five experiments that had used microdissected epithelial cells from colonic crypts. We considered these to be controls for the normal tissue samples because of the homogeneous cell type, whereas the normal tissue samples had detectable levels of stromal markers (*e.g.* vimentin). Hence genes with an average expression value less than 40 (below the detection limit of quantitative PCR) across the crypt samples were considered unexpressed in the epithelium layer and removed from consideration during scoring.

Mutual Information—MI is a concept from information theory used to measure the dependence of two random, discrete variables, say X and Y, based on their joint and marginal distributions. A high MI score ($0 \leq \text{MI} \leq 1$) indicates that X and Y are non-randomly associated to each other, whereas in the limit an MI value of 0 indicates that the two variables are statistically independent. To apply the concept to our problem, we retrieved the mRNA expression for every gene product in each of the four networks and used these data to populate two distributions of network activity values, one for a set of normal samples (*X*) and one for a set of stage D samples (*Y*). With these two distributions we were able to calculate an MI score between normal and stage D for each network and also use it as an optimization metric to search within the network for combinations of proteins (signatures) that would maximize this score. Intuitively a high MI score would indicate that the corresponding proteins non-randomly associate between normal and stage D, inferring their functional importance in the network in late stage cancer and suggesting new experiments for elucidating mechanisms of tumorigenesis.

The raw mRNA expression values in each subnetwork were first normalized by subtracting the population mean and dividing that difference by the population standard deviation. Next a subnetwork activity score was determined for each sample (column) in each network by summing the corresponding normalized mRNA expression values (rows). The activity score across phenotypes is a random, continuous variable that we discretized using a binning procedure. The number of bins (Fig. 5, *step 3*) was determined by $\log_2$(number of samples) + 1, which is Sturge's rule, in the same manner as described by Ideker and co-workers (15). The range of the bins was determined from the range of the normalized expression values plus or minus a small adjustment to ensure all values fell within a bin. Finally the marginal and joint distributions were determined for the two phenotypes being compared (normal and stage D), and the mutual information value was computed (Fig. 5, *steps 4* and *5*). Note that the example in Fig. 5 indicates that network activity values were calculated over every patient sample (column), which as a matter of course we did do, but we only calculated the MI between normal and stage D consistent with the proteomics comparison. The other values, however, were useful for testing hypothesis 2 (H2; see Fig. 6).

Significance Testing—To test the hypothesis that the genes in a given subnetwork were not significantly discriminative of phenotype compared with a random selection of *n* genes (where *n* equals the number of genes in the subnetwork being measured), we randomly selected 1,000,000 combinations of *n* genes from the decoy database to create the null distribution and then evaluated the actual MI score of the subnetwork on the cumulative distribution function (CDF). To test the second hypothesis, which is that the genes in our network do not associate with a particular phenotype, we permuted the phenotypes (columns) of the relevant array 100,000 times and evaluated significance in the same way. The evaluation on the CDF is expressed as a percent value. A value of, say, 95% indicates there is a 5% chance ($p = 0.05$) of observing a higher MI value, assuming the null hypothesis is true. The programs required to import the expression data, organize it for analysis, visualize it, and perform the scoring and optimization search as well as the hypothesis testing were written using Matlab and are available on request.

Label-free Mass Spectrometry: Protein -Fold Change Determination

Sample Preparation—50 μg of total protein derived from colonic tissue lysate was precipitated with acetone (−20 °C) at −80 °C for 20 min followed by centrifugation at 12,000 rpm for 10 min at 4 °C. The samples were then dried, rebuffered in 20 μl of 0.2% ProteasMax surfactant (Promega Corp., Madison, WI), and gently shaken for 30 min. The buffer volume was then increased to 93.5 μl with 50 mM ammonium bicarbonate and incubated with 1 μl of 0.5 M DTT for 20 min at 56 °C followed by incubation with 2.7 μl of 0.55 M iodoacetamide for 15 min in the dark. 1 μl of 1.0% ProteasMax was added to the buffer followed by 1.8 μl of trypsin (Promega Corp.) that had been dissolved in 50 mM ammonium bicarbonate to a final concentration of 1 μg/μl. Digestion was carried out for 3 h at 37 °C. Peptides were concentrated on a 100-μl C_{18} UltraMicroTip column (Net Group, Inc.), eluted in 20 μl of 0.1% formic acid in 60% acetonitrile, then diluted with UltraPure water to a final concentration of 500 ng/μl, and stored at −80 °C.

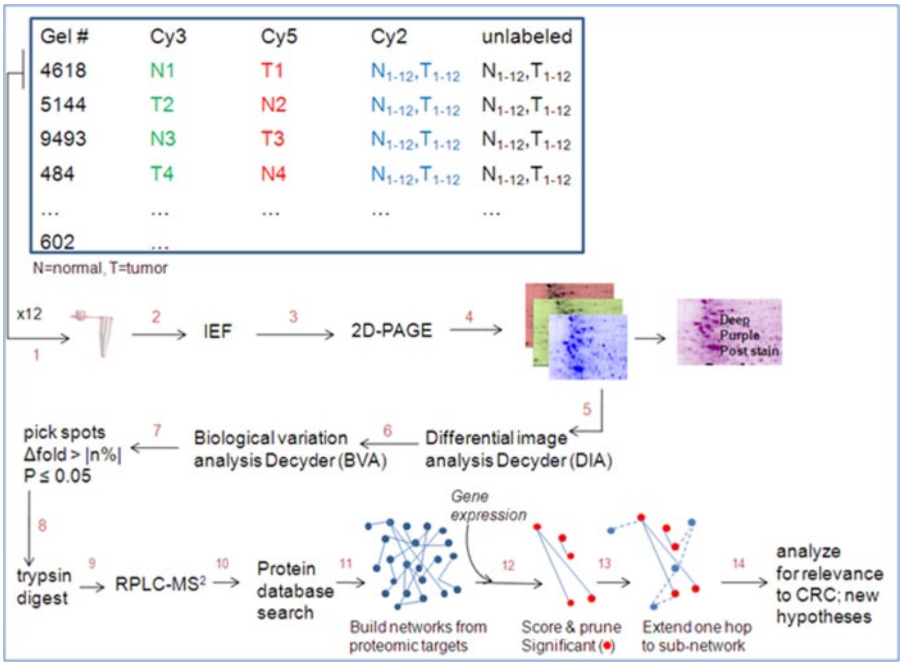
LC-MS/MS—Analyses were performed on an LC Packings/Dionex Ultimate 3000 HPLC-Orbitrap XL (Finnigan, San Jose, CA) system. The HPLC system is equipped with two independent ternary gradient pumps suitable for high throughput dual column parallel HPLC mode applications. A standard injection volume of 10 μl was used for all the samples, giving a total of 1 μg of digest on the column. The data collection method incorporated a 30,000 resolution Orbitrap full scan in the FT mode using profile mode data collection followed by data-dependent mode MS^2 acquisition of the top five precursors from each full scan (centroid mode). CID mode fragmentation was chosen for generating the MS^2 spectra in the linear ion trap with a standardized value of 30% normalized collision energy being chosen for the fragmentation of the peptides. The LC method included a slow, 95-min acetonitrile ramp from an initial 4.8% until a final composition of 50.2% was achieved. Further high organic elution was performed (5 min) to complete the elution of peptides from the analytical column followed by equilibration of the column for succeeding analyses.

Analysis—Raw files were searched by Mascot against the IPI_Human database (ipi.HUMAN.v3.28.fasta). For each protein of interest one tryptic peptide with tandem MS evidence in at least one of six replicate samples was used to measure the relative expression change between normal and tumor. Peptide abundance was determined by the area under the elution curve that was extracted from the total chromatogram using a mass window adequate to capture all isotopes of the observed monoisotopic mass (≤10 ppm). The curves were smoothed by an 11-point Gaussian filter and base line-subtracted using the Xcalibur software (Thermo Electron Corp.). -Fold change between normal and tumor was calculated as the ratio of the integrated curve areas. Three replicate runs of a single sample pair (patient 507, normal/tumor) were used to estimate technical variance. The coefficient of variation for -fold change ranged from 6 to 39%. The -fold change for IGFBP3 was determined by densitometry from Western blot analysis.

RESULTS AND DISCUSSION

2D DIGE Discovery Proteomics

The features of our cohort and the overall design of our proteomics experiments are shown in Fig. 1. Twenty-four tissue samples (12 normal and 12 tumor, each pair from the same patient) were prepared using standard procedures (see "Experimental Procedures"). The tissue samples had been vetted by a pathologist to establish tumor grade. The experimental design involved alternately labeling tumor and control samples with Cy3 and Cy5 dyes, whereas Cy2 was used to label a 50-μg pooled fraction that served as an internal standard for each gel. The usefulness of this standard cannot be overemphasized as it assists in providing a confident assessment of real biological variation by controlling for variance in protein loading (16). Each tripartite sample was first separated by IEF over a broad pH range (3–10) and then by molecular weight using 12.5% homogenous SDS-polyacrylamide gels. All the samples were labeled and separated simultaneously under identical conditions to minimize experimental variation



Gel #	Cy3	Cy5	Cy2	unlabeled
4618	N1	T1	N_{1-12},T_{1-12}	N_{1-12},T_{1-12}
5144	T2	N2	N_{1-12},T_{1-12}	N_{1-12},T_{1-12}
9493	N3	T3	N_{1-12},T_{1-12}	N_{1-12},T_{1-12}
484	T4	N4	N_{1-12},T_{1-12}	N_{1-12},T_{1-12}
...	...	...	...	...
602	...			

N=normal, T=tumor

x12
1
2
IEF
3
2D-PAGE
4
Deep Purple Post stain
5
Differential image analysis Decyder (DIA)
6
Biological variation analysis Decyder (BVA)
7
pick spots
Δfold > |n%|
P ≤ 0.05
8
trypsin digest
9
RPLC-MS²
10
Protein database search
11
Build networks from proteomic targets
Gene expression
12
Score & prune Significant (●)
13
Extend one hop to sub-network
14
analyze for relevance to CRC; new hypotheses

FIG. 1. **Experimental design.** For each patient, a tripartite sample of normal (*N*), tumor (*T*), and pooled control were mixed (*1*); unlabeled samples were run on separate gels for poststaining or mixed in the analytical gels (see "Experimental Procedures"). Samples were separated by isoelectric focusing (*2*) and then by molecular weight (*3*), and each fluorophore was imaged independently (*4*). Using the DeCyder software, spots were matched on an intragel basis with differential image analysis (*DIA*) (*5*) and on an intergel basis with biological variation analysis (*BVA*) to assess biological variation (*6*). Significant spots (Δfold dependent on statistical power) were selected for robotic excision (*7*) and digested by trypsin (*8*), and the peptides were separated by reverse-phase (*RP*) chromatography and detected by tandem mass spectrometry (*9*). MS^2 spectra were searched using SEQUEST (*10*), and identified proteins were imported to MetaCore to search for relevant networks (*11*). Significant protein signatures are scored by mutual information using gene expression profiles (*12*); signatures are extended out one hop to infer functional relevance (*13*). Resultant subnetworks were analyzed for biological relevance to CRC and new hypothesis generation (*14*).

(see "Experimental Procedures"). Each gel yields three images (Fig. 1, *step 4*), and along with the poststained gel used for spot excision, a total of 37 images were imported to the DeCyder software (GE Healthcare) for differential image analysis (*i.e.* spot matching) followed by statistical analysis of biological variation. Fig. 2 is a Cy5 image representative of a typical analytical gel (patient 5144) indicating the significant spots matched. In total for this experiment 58 spots were identified as significantly ($p \leq 0.05$, mean fold change $>50\%$) changing between normal and tumor. For the majority of these spots DeCyder was able to detect a match on greater than 30 of the images. In no case did that number fall below 20 or fail to be matched on the poststained gel used for spot excision. The spots were robotically excised from gel, digested by trypsin overnight, and submitted to reverse-phase LC-MS^2 followed by database search. Twenty-three spots were confidently ($p \leq 0.001$, peptide and protein) identified by database search (Table I and Fig. 2, annotated). Thirteen proteins were up-regulated in cancer, and seven were down-regulated.

The IEF range chosen for this experiment resulted in spot overlay in a number of regions of interest. We anticipated that we could improve separation and focus more spots by using an IEF range of pI 4–7. Additionally using a measure of spot variance from the first experiment, we found we only needed six sample pairs to capture -fold changes greater than ±30% while maintaining statistical power of 0.8 (see "Statistical Power Analysis" under "Experimental Procedures"). Accordingly we performed a second experiment using a subset of six sample pairs from the first experiment (numbers 145, 321, 362, 468, 480, and 602). The protein fractions were thawed and cleaned with a kit (2D-Cleanup kit, GE Healthcare) to remove impurities known to interfere with proper separation. The protein concentration was redetermined as before by colorimetric assay. We used even more stringent criteria in the image analysis phase as compared with experiment 1. Spots not only had to satisfy the same statistical criteria, but additionally, for a spot to be considered for picking, it had to have been matched by DeCyder on *every one* of the 19 gel images. Using these criteria we identified 150 significantly ($p \leq 0.05$, mean -fold change $>30\%$) changing spots (Fig. 3). Activating the false discovery rate filter in DeCyder (based on the method of Benjamini and Hochberg (17)), over 40 of these spots retained their significance ($q \leq 0.05$).The indicated spots were excised from the gel, digested by trypsin, and submitted to reverse-phase LC-MS^2 followed by database search. Of 150 spots, we confidently identified 67 proteins. Thirty-five spots were up-regulated in cancer, and 32 were down-regulated (Table II), indicative of a lack of bias toward up- or down-regulated proteins.

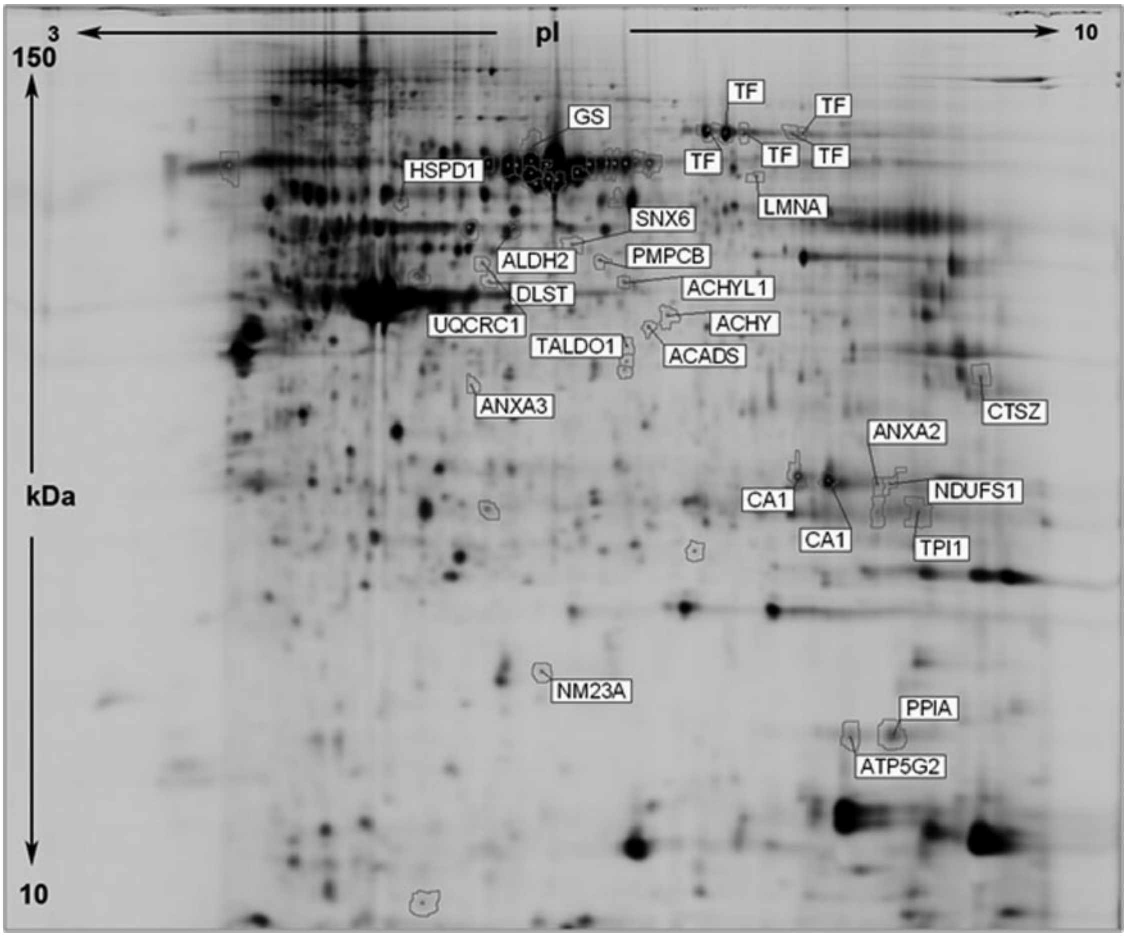


FIG. 2. **Representative gel from experiment 1 (5144).** *Polygons* indicate spots significantly changing between normal and cancer as determined by DeCyder. Spots identified by mass spectrometry are labeled. See Table I.

Highly Significant Late Stage CRC Signatures

As stated above, our guiding hypothesis was that the differentially expressed proteins ("targets") found in our experiments represented nodes upstream and/or downstream of other nodes in one or more functional subnetworks that may be dysregulated in stage D colon cancer. Accordingly we used the set of unique targets from both experiments ($n = 67$) to seed a search for functional subnetworks. A detailed description of the protein interaction database we searched and the subnetwork construction algorithm we used is provided under "Experimental Procedures." The search returned 13 subnetworks, each of which contained a variable number of between one and 12 of the seed targets. We limited our attention to four subnetworks judged most significant by a combination of *p* value and zScore. One of these subnetworks is annotated in Fig. 4 with a breakdown of the most significant gene ontological processes, their percent representation in the subnetwork, the *p* value, zScore, and subnetwork size, *i.e.* the total number of gene products it contains. The remaining three are provided in supplemental Data S4.

To test our hypothesis that our 67 proteomic targets were significant for late stage CRC, we implemented a quantitative method to score the subnetworks. Fig. 5 outlines our approach. A more detailed explanation of the scoring procedure is provided under "Experimental Procedures." Briefly we obtained mRNA expression data for every gene product in each of the four subnetworks from a set of unpublished microarray experiments (Affymetrix) performed on a large cohort of clinical tissue samples of varying CRC stage (Fig. 5, *step 1*). With these data we computed an activity value for each subnetwork over each normal experiment ($n = 16$) and each stage D experiment ($n = 50$) (Fig. 5, *step 2*). The activity values represent a continuous variable and for the purpose of computing MI need to be discretized by a binning procedure (Fig. 5, *step 3*). With discrete values we computed the relevant distributions (Fig. 5, *step 4*) and then the MI between normal and stage D for each subnetwork. The scores are shown in Table III (last column).

There are two null hypotheses relevant to evaluate the significance of the MI score. The first, which we will call H1, states that genes in a subnetwork are not discriminative of

TABLE I

List of unique proteins (18) from a total of 20 identified by experiment 1, pI 3–10

When a protein was identified at multiple spots on the gel, the -fold change here represents the average value.

Gene	Protein	NCBI gi|number	-Fold change	Molecular weight	Theoretical pI
ACADS	Acyl-coenzyme A dehydrogenase, C-2 to C-3 short chain	gi|19684166	−1.82	44,299.8	7.96
AHCY	*S*-Adenosylhomocysteine hydrolase	gi|9951915	1.51	47,799.3	5.9
ALDH2	Mitochondrial aldehyde dehydrogenase 2	gi|48256839	1.56	56,318.6	6.37
ANXA2	Annexin A2	gi|16306978	−2.46	38,594	7.77
ANXA3	Annexin III	gi|12654115	1.93	36,452.7	5.69
CA1	Carbonic anhydrase I	gi|4502517	−2.8	28,853.4	6.67
DLST	Dihydrolipoamide succinyltransferase	gi|643589	2.66	48,555	8.89
HSPD1	Heat shock protein 60	gi|77702086	2.15	61,175.5	5.59
LMNA	Lamin A/C isoform 3	gi|27436948	1.55	65,208	8.5
NDUFS1	*Homo sapiens* similar to zinc finger protein (LOC147947), mRNA	gi|18490405	−1.69	79,417.5	5.84
NM23A	Nucleoside-diphosphate kinase A	gi|35068	1.79	20,399.3	7.07
PMPCB	Mitochondrial processing peptidase β subunit precursor	gi|94538354	1.75	54,332.5	6.38
PPIA	Predicted: similar to peptidylprolyl isomerase A isoform 1	gi|89058333	1.57	24,502	7.11
SERPINF1	Pigment epithelium-derived factor	gi|15217079	−1.39	46,314	5.95
SNX6	Sorting nexin 6 isoform b	gi|88703041	1.52	47,775.3	5.99
TALDO1	Transaldolase 1	gi|5803187	1.24	37,517.5	6.38
TF	Transferrin	gi|37747855	−1.66	77,030.6	6.86
TPI1	Predicted: similar to triose-phosphate isomerase (TIM) (triose-phosphate isomerase) isoform 8	gi|88942747	−1.85	26,926	8.1

the disease phenotype compared with a random set of genes. For example, if the subnetwork contained 10 genes, then any 10 genes taken at random would produce an MI score at least as good as the real subnetwork of 10 genes under the null hypothesis. The second, call it H2, is subtly different; it states that the expression levels of the genes in the subnetwork do not associate with a particular phenotype. For example, if the network contains 10 genes that produce a high MI score between normal and stage D, then under the null hypothesis scores at least as high will be found for random permutations of phenotypes, *i.e.* by disrupting the real association between patient and gene expression.

From the microarray we obtained mRNA expression data for 1000 random genes ("decoys") and ensured that these genes had no overlap with any of the genes in the four subnetworks. A null distribution was estimated for testing H1 by evaluating an MI score for 1,000,000 combinations of *n* random genes selected from the decoy data set where *n* equals the size of the particular subnetwork being assessed. The null distribution for H2 was estimated from 100,000 permutations of the phenotypes (columns) in the two-dimensional array representing a subnetwork. Significance was then determined by evaluating the MI score on the CDF of the respective null distribution. 1 − CDF indicates the probability of finding a higher MI score. Probability values of 1% or less were considered to be significant.

By this measure neither the null hypothesis for H1 nor that for H2 could be rejected for any of the four subnetworks, *i.e.* when *all* the gene products in the subnetwork were used to compute its activity value (Table III, last column). We reasoned that this result could be attributed to how the subnetworks were built and scored. Although all four subnetworks were discovered by proteomics profiling and were judged statistically significant, the individual interactions in a subnetwork are nevertheless based in large part on a diverse set of experiments performed *in vitro* and *in vivo* in a variety of different tissues. Consequently although many of the subnetworks are indicated to be active in colonic tissue, they are not necessarily important to a metastatic cancer phenotype. Also we observed that the subnetworks, in terms of the nodes and edges they contain, were sensitive to the parameters used to search for them, even including the version of the database software. This results in a certain degree of arbitrariness in the overall topology of the subnetwork. Given these caveats the statistically insignificant MI scores were not very surprising.

However, we further hypothesized that the activity of specific protein combinations within the subnetwork(s) would be highly discriminative of disease. Thus, we performed an exhaustive combinatorial search over each subnetwork for combinations that would maximize the MI score between control and stage D. We did this for up to six combinations (readily accomplished on a conventional

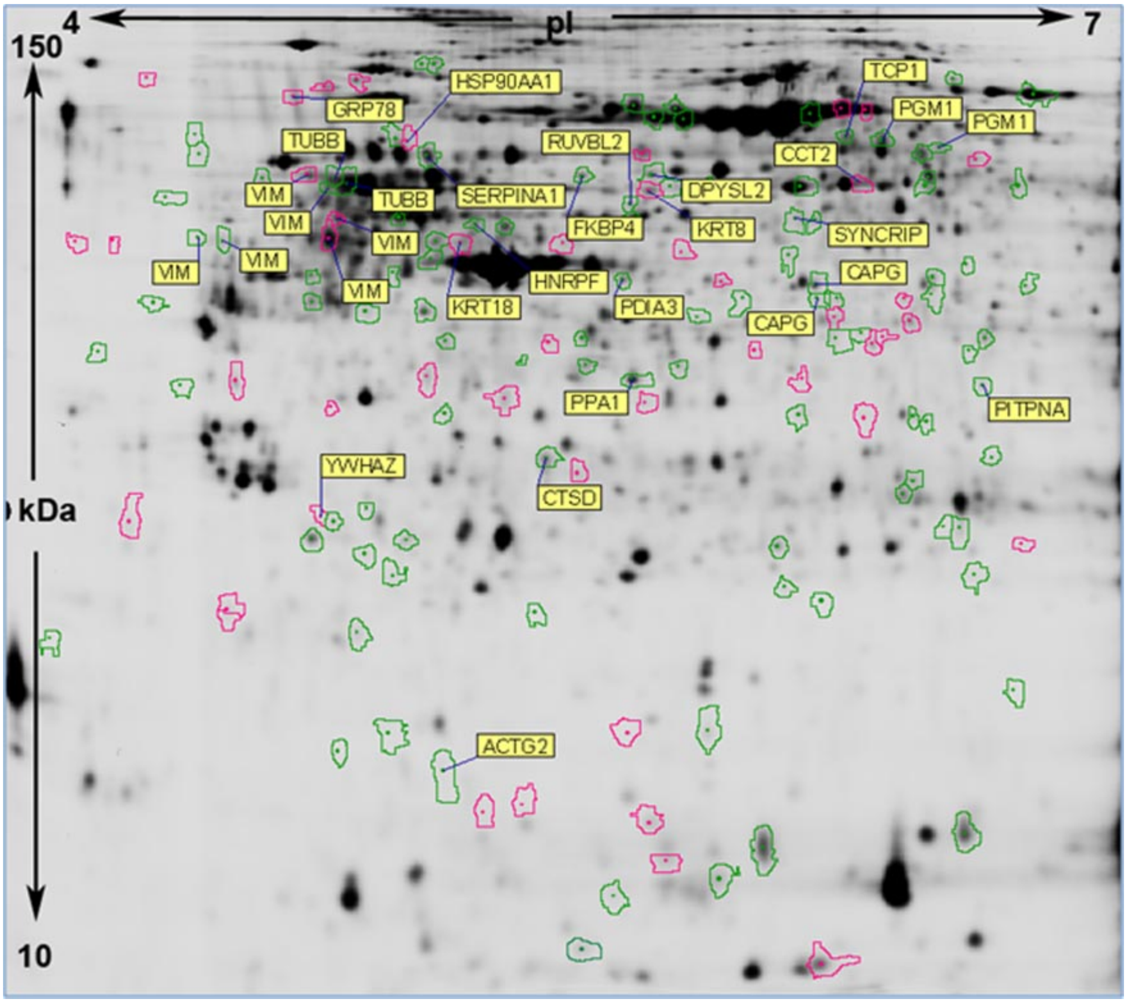


FIG. 3. **Representative gel from experiment 2.** *Polygons* indicate spots significantly changing between normal and cancer as determined by DeCyder. *Magenta polygons* indicate that the spot passed a multiple comparison filter test (false discovery rate) in DeCyder. Labeled spots were identified by mass spectrometry and appeared in one or more of four significant MetaCore networks. See Table II.

desktop computer). For each subnetwork except one (subnetwork 2), the MI score steadily increased with combination number (Table III). MI scores for these specific combinations (signatures) were much higher than those for any of the subnetworks taken as a whole, and more importantly, in each case the MI score was highly significant with respect to both null hypotheses, H1 and H2 (Fig. 6, compare with MI in column labeled "signature 6" in Table III). Notably the top scoring signatures included proteins for which we had independently found direct proteomics evidence (*e.g. CCT2*, *HSP90AB1*, *SERPINA1*, and *CapG*) as did certain other signatures of fewer combinations. Fig. 7 highlights the gene products (*gray*) from each subnetwork participating in signature 6. To extend the potential functional importance of these signatures, we added to each directed graph those proteins that were one hop away from the signature proteins, *i.e.* those from the corresponding parent subnetwork immediately up- or downstream. We briefly discuss the proteins and functional interactions of these expanded subnetworks in more detail in the following section. Additionally similar to the conclusion of Ideker and co-workers (15), we found that the signature genes did not cluster in a dendrogram computed using traditional distance metrics, *e.g.* Euclidean or Spearman (data not shown), and would likely have been overlooked by conventional gene classification techniques.

As mentioned above, with the exception of parent network 2, MI scores increased or were constant for successively larger combinations of proteins. Further we found that combinations of less than six proteins (signatures 1–5) were also significant when tested against the appropriate H1 and H2 null distributions (data not shown). Indeed the relevant distributions for smaller combinations of proteins had characteristics (*e.g.* mean and variance) similar to those for signature 6. If the proteins appearing in successively larger combinations were completely different this would suggest that our method was not sensitive to small variations in the underlying network activity patterns. However, this is not

TABLE II

List of unique proteins from a total of 67 identified by experiment 2, pI 4–7

Y (yes) or N (no) indicates whether the protein survived the false discovery rate (FDR) significance test. When a protein was identified at multiple spots on the gel, the -fold change in this table usually represents the average value.

Gene	Protein	NCBI gi|number	-Fold change	FDR	Molecular weight	Theoretical pI
ACTB	β-Actin	gi|4501885	3.54	Y	41,170	5.18
ACTG2	ACTG2	gi|49168516	3.26	N	41,898	5.2
ACTR3	ARP3 actin-related protein 3 homolog	gi|5031573	1.77	N	47,432	5.54
ALDH2	Mitochondrial aldehyde dehydrogenase 2	gi|48256839	2.17	N	56,420	6.67
ANXA4	Annexin IV	gi|4502105	1.43	Y	36,063	5.75
ANXA5	Annexin V	gi|49168528	1.99	Y	35,941	4.78
APOH	Apolipoprotein H (β_2-glycoprotein I)	gi|18089104	−1.52	N	38,273	7.84
ATP5B	ATP synthase, H^+-transporting, mitochondrial F_1 complex, β	gi|32189394	−1.38	N	56,525	5.14
CA1	Carbonic anhydrase I	gi|4502517	−1.77	N	28,853	6.67
CapG	Gelsolin-like capping protein	gi|63252913	1.69	N	38,475	5.79
CAPNS1	Calpain, small subunit 1	gi|40674605	2.68	N	28,212	4.82
CAPZA1	F-actin capping protein α-1 subunit	gi|5453597	−2.59	N	32,903	5.36
CCT2	Chaperonin containing TCP1, subunit 2	gi|5453603	−1.55	Y	57,453	6
CES1	Carboxylesterase 1 isoform c precursor	gi|68508957	−1.96	Y	62,354	6.15
COMT	Catechol-*O*-methyltransferase isoform S-COMT	gi|6466450	1.52	N	24,434	5.02
CTSD	Cathepsin D	gi|30584113	−1.47	N	44,637	6.1
CTSX	Preprocathepsin P	gi|3719219	−1.74	N	32,681	6.1
DPYSL2	Dihydropyrimidinase-like 2	gi|4503377	1.57	N	62,255	5.93
ECH1	Peroxisomal enoyl-coenzyme A hydratase-like protein	gi|70995211	−1.68	N	35,972	6.68
FGB	Fibrinogen, β chain preproprotein	gi|70906435	6.53	N	55,893	8.23
FGG	Fibrinogen γ-prime chain	gi|182440	1.77	N	51,464	5.19
FKBP4	FK506-binding protein 4	gi|4503729	1.59	N	51,773	5.22
HNRPF	Heterogeneous nuclear ribonucleoprotein F	gi|4826760	1.99	N	45,643	5.27
HNRPH1	Heterogeneous nuclear ribonucleoprotein H1	gi|5031753	−1.51	Y	49,199	5.86
HP	Haptoglobin	gi|4826762	−2.53	N	45,177	6.13
HPX	Hemopexin	gi|11321561	−1.39	N	51,643	6.6
HSP90	Heat shock protein gp96 precursor	gi|15010550	2.36	N	92,412	4.61
HSP90AA1	HSP90AA1 protein	gi|12654329	1.86	Y	64,350	4.96
HSP90AB1	HSP90AB1 protein	gi|39644662	1.57	N	74,769	4.91
HSPA5	Heat shock 70-kDa protein 5	gi|16507237	−1.86	Y	72,289	4.92
IMPDH2	IMP (inosine monophosphate) dehydrogenase 2	gi|15277480	−1.52	N	55,770	6.46
KRT18	KRT8 protein	gi|33875698	2.85	Y	55,788	5.49
KRT8	Keratin 8	gi|4504919	6.13	Y	53,672	5.38
KRT9	KRT9 protein	gi|113197968	1.52	N	48,057	4.7
LDHD	D-Lactate dehydrogenase isoform 2 precursor	gi|37595756	6.32	Y	52,112	6.02
MAPRE1	Microtubule-associated protein, RP/EB family, member 1	gi|6912494	2.05	N	29,981	4.87
MRLC2	Myosin regulatory light chain MRCL2	gi|15809016	1.71	Y	19,767	4.54
NNMT	Nicotinamide *N*-methyltransferase	gi|5453790	2.24	Y	29,556	5.46
OXCT	Succinyl-CoA:3-ketoacid-coenzyme A transferase 1, mitochondrial precursor	gi|48146215	−1.57	N	56,159	7.21
PDIA3	Protein-disulfide isomerase-associated 3 precursor	gi|21361657	−1.54	N	56,747	5.95
PDIA5	Protein-disulfide isomerase-related protein 5	gi|1710248	−1.38	N	46,171	4.81
PGAM1	Phosphoglycerate mutase 1 (brain)	gi|4505753	2.62	N	28,802	6.79
PGM1	Phosphoglucomutase 1	gi|21361621	−2.18	N	61,411	6.31
PITPNA	Phosphatidylinositol transfer protein, α	gi|5453908	−1.35	N	31,787	6.11
PKM2	PKM2 protein	gi|33870117	−1.76	Y	61,362	8.86
PMPCB	Mitochondrial processing peptidase subunit β, mitochondrial precursor	gi|40226469	1.84	N	53,475	6.3
PPA1	Inorganic pyrophosphatase	gi|33875891	1.67	N	35,449	5.92
RRBP1	Ribosome-binding protein 1	gi|110611220	−1.77	Y	108,590	5.33
RUVBL2	RuvB-like 2	gi|5730023	1.57	N	51,125	5.37
SELENBP1	Selenium-binding protein 1	gi|16306550	−1.89	N	52,357	5.9
SEPT2	SEPT2 protein	gi|23274163	1.74	N	42,659	6.4

TABLE II—*continued*

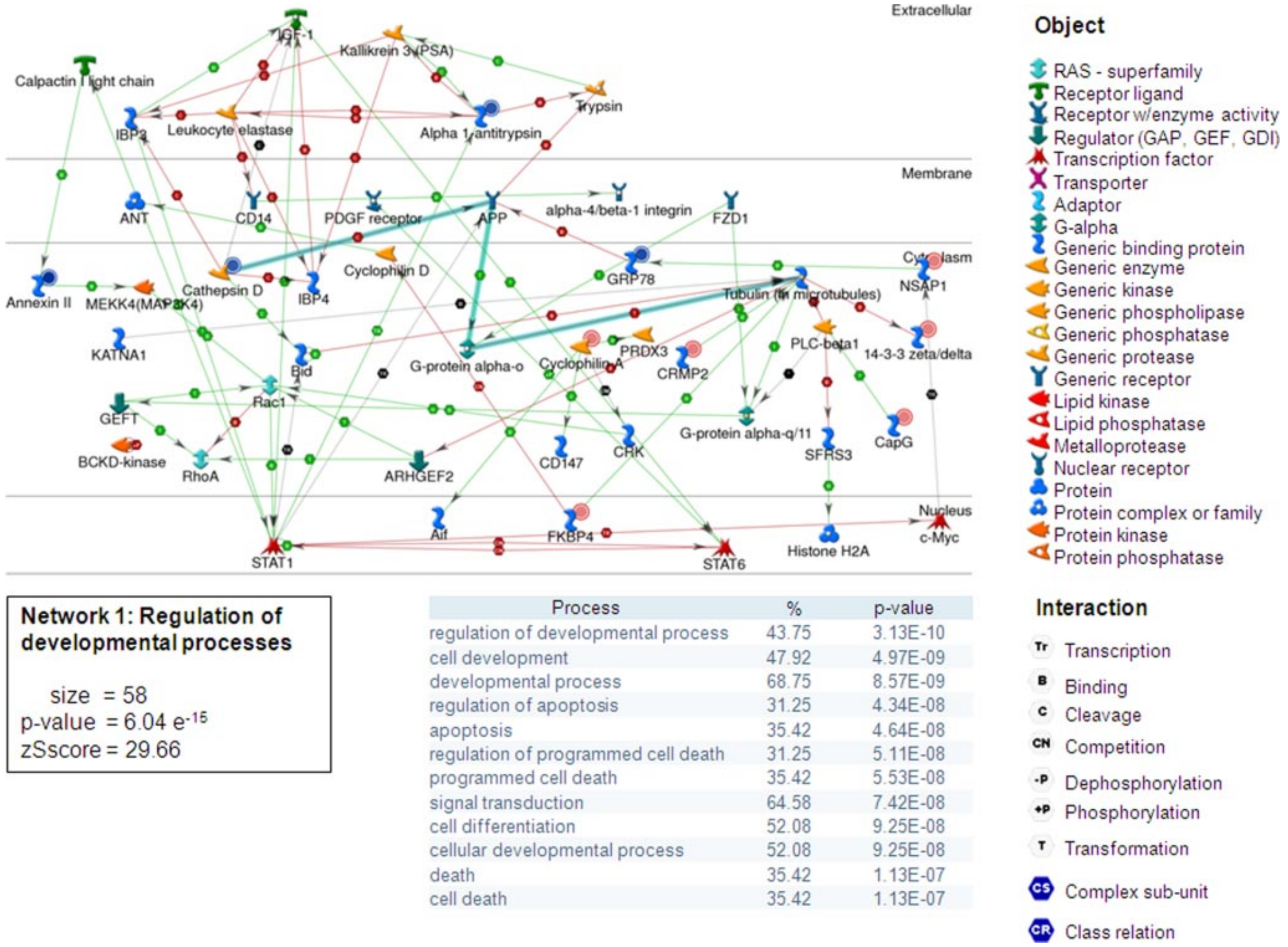
Gene	Protein	NCBI gi|number	-Fold change	FDR	Molecular weight	Theoretical pI
SERPINA1	Serine (or cysteine) proteinase inhibitor, clade A (α-1)	gi|50363219	−1.66	N	46,588	5.27
SERPINB2	Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin)	gi|62898301	1.55	Y	42,743	5.87
SERPINB6	SERPINB6 protein	gi|12655087	−1.98	N	42,563	5.1
SYNCRIP	Heterogeneous nuclear ribonucleoprotein Q	gi|33874520	1.84	N	46,715	5.81
TAGLN	Transgelin	gi|48255905	4.14	N	22,596	9.4
TALDO1	Transaldolase 1	gi|5803187	1.46	Y	37,517	6.38
TCP1	*H. sapiens* t-complex 1	gi|30584211	−2.02	N	60,419	5.74
TF	Transferrin	gi|4557871	−1.54	N	77,000	6.75
TPI1	Triose-phosphate isomerase 1	gi|17389815	−1.51	N	26,624	6.5
TUBB	Tubulin, β	gi|18088719	−1.75	N	49,641	4.9
TXNDC4	Thioredoxin domain-containing 4 (endoplasmic reticulum)	gi|52487191	2.55	N	46,900	5.01
TXNDC5	Thioredoxin domain-containing 5 isoform 2	gi|42794775	2.2	N	43,642	5.73
VIM	Vimentin	gi|47115317	−3.05	Y	53,548	4.94
YWHAZ	14-3-3 protein ζ/δ	gi|49119653	1.99	Y	29,928	4.57

what we observed. As indicated in Table III, for each network all signatures (1–6) frequently show the repeated contribution of either the same protein or proteins involved in the same complex of proteins, suggesting that the method is sensitive to the specific interactions of functionally similar subnetwork proteins. Even for the signatures derived from parent subnetwork 2, the contribution of proteins capable of phospholipase activity (*PLA2* and *PRDX6*) appears in five of six signatures.

Biological Relevance to CRC of Signature Proteins in Extended Subnetworks

It merits emphasis that each of the most significant extended subnetworks (Fig. 7) contained targets for which we had direct proteomics evidence (*CCT2*, *HSP90AB1*, *SERPINA1*, and *CapG*), indicating that gene products significant by their contribution to MI maintain their significance at the level of the proteome in late stage CRC. The most significant targets (highlighted *gray*) are generally classified according to those with a known role in CRC or a role in other human cancers and those with no known role in cancer. The fact that we found significant genes with a known role in CRC can be understood as a positive control of our analytical method.

Genes with a Role in CRC—IGFBP3 (also known as *IBP3*) is an insulin-like growth factor-binding protein that was recently identified to cause apoptosis in a tumor necrosis factor-related apoptosis-inducing ligand (TRAIL)-mediated fashion in relevant CRC cells (18). Additionally a large association study found that paired polymorphisms in *IGFBP3* and its substrate predicted a significant increase in risk for CRC (19). The integrin family of proteins has been well studied in CRC. They are generally responsible for cell-cell and cell-matrix adhesion. Loss of expression of certain subunits in this family has been associated with increased neoplastic transformation in colonic epithelium, and the specific loss of β1 (*ITGB1*) chains was associated with benign to malignant transformations (20). Notably integrins are active as heterodimers, and although only *ITGB1* contributed to the significant MI score, the subnetwork indicates that the dimer *ITGB1*/*ITGA4* is of particular interest. *IFITM1* (*IFI17*) is a member of the family of interferon-inducible transmembrane proteins. A recent study (21) proposed it as a possible marker for human colorectal tumors. The subnetwork also revealed it to be regulated at the level of transcription by the PBAF complex, certain members of which we also found significant. *SMARCA4* (also known as *BRG-1*) plays a key role in the chromatin-remodeling complex in mammals. One study (22) showed *in vivo* evidence that BRG-1 interacts with β-catenin and induces the transcription of T-cell factor (TCF) target genes. Mutant forms of *BRG-1* lacking ATPase activity disrupted this induction. TCF target gene activation is the final consequence of the WNT signaling pathway, mutations in which are well known to cause tumorigenesis in the human colon. The role of platelet-derived growth factor receptor (*PDGFR*) has been well studied in CRC along with other receptors capable of tyrosine kinase activity and downstream signaling. One recent study (23) found a significant association between the stromal expression of the B subunit (*PDGFRB*) and the metastatic potential of CRC tumors. In addition to being overexpressed in a number of human cancers, casein kinase II (*CSNK2A2*) in CRC suppresses apoptosis by desensitizing cells to TRAIL in a caspase-dependent manner but independent of NF-$\kappa\beta$ (24). It has also been shown to promote survival of colon cancer cells by increasing the expression of survivin via the canonical transcription pathway hyperactive in CRC (TCF/LEF (lymphoid enhancer binding factor)) (25). *PLA2G12A* is a member of the family of secreted phospholipases, many of which display distinct patterns of expression in adenocarcinomas (26).



Network 1: Regulation of developmental processes

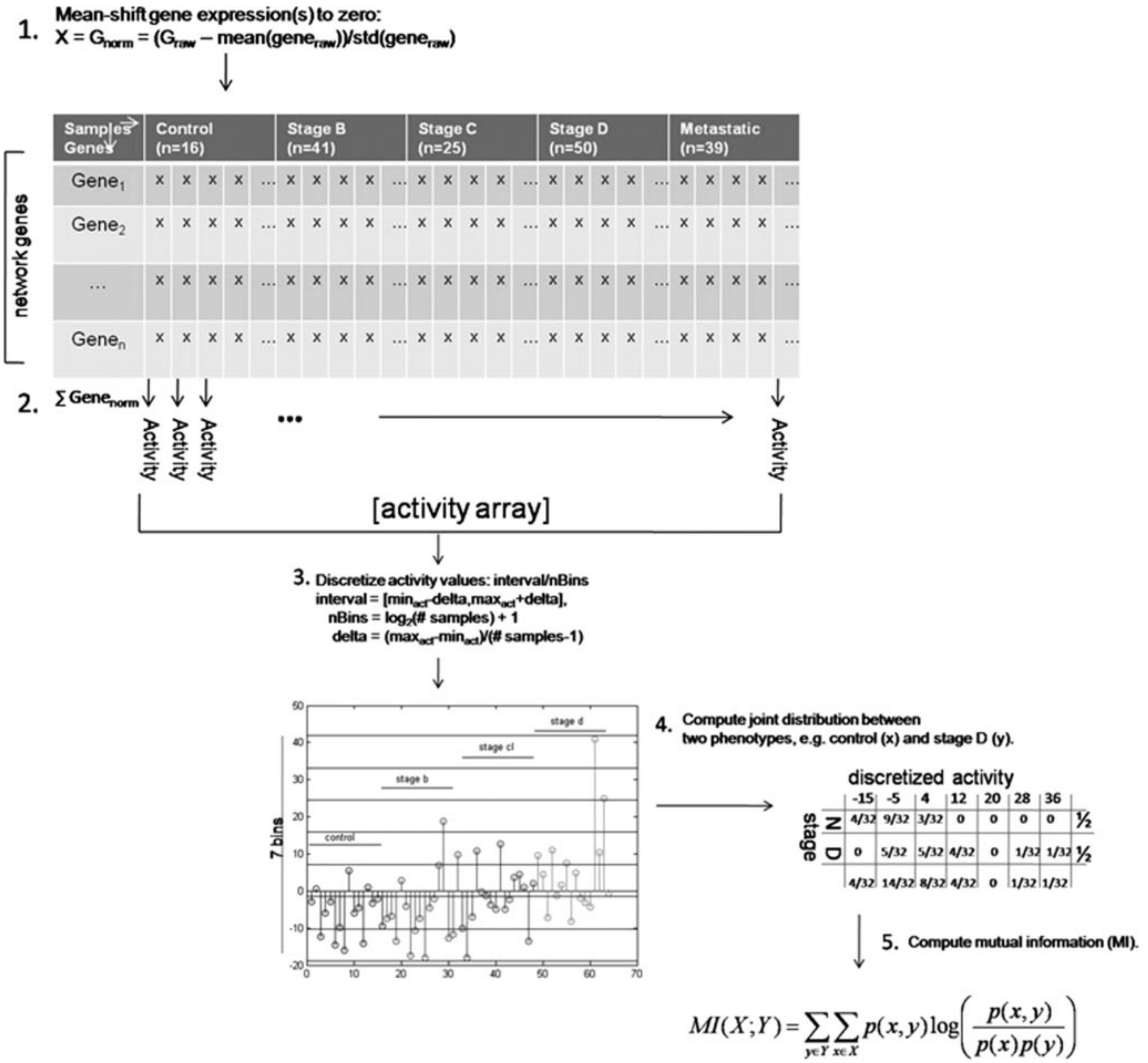
size = 58
p-value = 6.04 e^{-15}
zSscore = 29.66

Process	%	p-value
regulation of developmental process	43.75	3.13E-10
cell development	47.92	4.97E-09
developmental process	68.75	8.57E-09
regulation of apoptosis	31.25	4.34E-08
apoptosis	35.42	4.64E-08
regulation of programmed cell death	31.25	5.11E-08
programmed cell death	35.42	5.53E-08
signal transduction	64.58	7.42E-08
cell differentiation	52.08	9.25E-08
cellular developmental process	52.08	9.25E-08
death	35.42	1.13E-07
cell death	35.42	1.13E-07

FIG. 4. **MetaCore subnetwork.** Shown is a characteristic example of one of four significant MetaCore protein interaction subnetworks returned by a search seeded by significant proteomic targets: subnetwork 1, regulation of developmental processes. Interaction effects are positive (*green*), negative (*red*), and unspecified (*black*). *Red* and *blue circles beside* certain objects indicate that the protein was identified by proteomics, either up-regulated in cancer (*red*) or down-regulated in cancer (*blue*). *Size* indicates the total number of gene products used for scoring by mutual information. Similar details for each of the other three subnetworks chosen for scoring are provided in supplemental Data S4.

Genes with a Role in Other Human Cancers—CapG, a gelsolin-like capping protein, has been identified as a possible tumor suppressor gene (27), although our proteomics screen revealed it to be up-regulated in cancer in agreement with the mRNA expression. A closer look at this study revealed that the authors had measured a near complete loss of the *CapG* protein in a variety of primary human cancer tissues but not colon tissue. Our evidence that the *CapG* message and protein are up-regulated in CRC indicates that it may have oncogenic activity in the colon. Further we actually identified *CapG* at two closely spaced but distinct spots on the gel, suggesting that post-translational modification may be important to its activity. The human gene *PLK1* (or *PLK*), a serine/threonine protein kinase, was characterized many years ago (28), and its expression was found to strongly correlate with the mitotic activity of a variety of tumor cell lines, including those derived from human colon. Notably the study found that *PLK1* was not expressed in a variety of the normal human tissues with the exception of normal colon tissue. More recently, *PLK1* was found to be overexpressed in primary CRC tumors (29), identified as a prognostic factor for CRC (30), and when knocked down or inhibited in human adenocarcinoma cells (RKO) lead to dramatic mitotic arrest (31), thus showing promise as a possible drug target. Lastly driver mutations in *PLK1* are not unknown as was recently revealed by a large screen for somatic mutations on over 500 protein kinases covering a large cohort of human cancers (32). *RPS2* is a gene encoding a ribosomal protein in the small 40 S subunit. *RPS2* was found by a proteomics screen to be a novel kinase substrate differentially phosphorylated in breast cancer development (33).

Genes with No Known Role in Human Cancer—A literature search of PubMed revealed little to nothing about the role of


1. Mean-shift gene expression(s) to zero:
X = G_norm = (G_raw – mean(gene_raw))/std(gene_raw)
Samples
Genes
Control (n=16)
Stage B (n=41)
Stage C (n=25)
Stage D (n=50)
Metastatic (n=39)
network genes
Gene1
Gene2
...
Genen
x
2. Σ Gene_norm
Activity
[activity array]
3. Discretize activity values: interval/nBins
interval = [min_act-delta,max_act+delta],
nBins = log2(# samples) + 1
delta = (max_act-min_act)/(# samples-1)
7 bins
control
stage b
stage cl
stage d
50
40
30
20
10
0
-10
-20
0
10
20
30
40
50
60
70
4. Compute joint distribution between two phenotypes, e.g. control (x) and stage D (y).
discretized activity
stage
5. Compute mutual information (MI).

	-15	-5	4	12	20	28	36	
N	4/32	9/32	3/32	0	0	0	0	½
D	0	5/32	5/32	4/32	0	1/32	1/32	½
	4/32	14/32	8/32	4/32	0	1/32	1/32	

$$MI(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

FIG. 5. **Flow chart showing steps required to compute MI.** Gene expression values (*X*s, *rows*) were mean-shifted to 0 across samples (*columns*) (*1*). Normalized values were then summed to produce an activity value for each sample (*2*); activity scores are continuous and need to be assigned into discrete bins for an MI calculation (*3*). A joint distribution matrix is calculated between two sets of samples, *e.g.* normal (*N*) and stage D (*4*). MI is calculated as shown where *p*(*x*) is the marginal distribution of normal activity, *p*(*y*) is the marginal distribution of stage D activity, and *p*(*x*,*y*) is the joint distribution of *x* and *y* (*5*).

the remaining significant subnetwork targets in human cancer (*HSP90AB1*, *HIST1H2AB*, *TUBA4A*, *TUBB3*, *GNA12*, *TRAP1*, *DYNLT3*, *CCT3*, *CCT5*, *CCT7*, and *POLR2D*). Guided by the evaluation of interactions on the subnetwork along with select proteomics evidence, these targets may merit follow-on experiments to discover their role, if any, in late stage CRC. For example, the chaperone containing t-complex proteins (CCTs) play a role in protein folding in eukaryotes and are widely expressed in the cytosol. One study did find a significant elevation of the CCT transcript in human colon carcinomas and validated the change in protein expression by immunohistochemistry, whereas our proteomics screen revealed it to be down-regulated in the cancer tissue. Notably that study had not vetted the samples for tumor stage, highlighting the importance of stage-specific studies. Additionally *CCT3* and *CCT5* were identified by microarray analysis to be significantly differentially expressed in the epithelium of other human cancer tissues (esophageal, breast, ovarian, and lung), but their functional role in cancer, CRC in particular, is unknown.

TABLE III
The MI scores for each signature

Signature 1 represents the single best protein by the measure of MI, signature 2 represents the highest scoring combination of two, signature 3 represents the highest scoring combination of three, etc. MI values of the corresponding parent subnetwork (Fig. 4 and supplemental Data S4) appear in the last column. p values are included for signature 6 and the whole parent subnetwork: p_{H1}, probability of achieving a higher MI value under null hypothesis H1; p_{H2}, probability of achieving a higher MI value under null hypothesis H2. Genes in bold font indicate proteins with direct proteomics evidence.

(MetaCore parent network no.)	Signature 1	Signature 2	Signature 3	Signature 4	Signature 5	Signature 6 p_{H1} p_{H2}	Whole network p_{H1} p_{H2}
MI(1)	0.4116	0.4326	0.4525	0.4545	0.4820	**0.4981** $p_{H1} = 0.0004$ $p_{H2} \ll 0.0001$	0.1774 $p_{H1} = 0.73$ $p_{H2} = 0.63$
Genes(1)	*PDGFRB*	*TUBA4A* *TUBB3*	*H2AFX* *TUBA1A* *TUBB3*	*IGFBP4* *PPID* *TUBA4A* *TUBB3*	*IGFBP4* *PPID* *TUBA4A* *TUBB3* *HIST1H2AB*	***CapG*** *HIST1H2AB* *IGFBP3* *ITGB1* *TUBA4A* *TUBB3*	Fig. 4, subnetwork 1
MI(2)	0.4971	0.4981	0.4713	0.4786	0.4530	**0.4628** $p_{H1} = 0.0009$ $p_{H2} \ll 0.0001$	0.1668 $p_{H1} = 0.92$ $p_{H2} = 0.82$
Genes(2)	*PRDX6*	*PLA2G10* *PRDX6*	*FOS* *PLA2G10* *PRDX6*	*CSNK2A2* ***HSP90AA1*** *PLK1* *RB1*	*FOS* *PLA2G10* *PLA2G4A* *PLA2G6* *PRDX6*	*CSNK2A2* *GNA12* *PDGFRB* *PLA2G12A* *PLK1* ***SERPINA1***	Supplemental Data S4, subnetwork 2
MI(3)	0.4971	0.5171	0.5717	0.5717	0.6063	**0.6063** $p_{H1} = 0.0007$ $p_{H2} \ll 0.0001$	0.1879 $p_{H1} = 0.42$ $p_{H2} = 0.44$
Genes(3)	*PRDX6*	*CCT3* *TRAP1*	***PPAI*** *TRAP1* *TUBA1A*	***HSP90AA1*** ***PPA1*** *TRAP1* *TUBA1A*	***CCT2*** *CCT6A* *CCT7* *SMARC4A* *TRAP1*	*CCT3* *CCT5* *CCT7* *DYNLT3* *PLK1* *TRAP1*	Supplemental Data S4, subnetwork 3
MI(4)	0.4000	0.4552	0.4983	0.5057	0.5462	**0.5398** $p_{H1} = 0.0002$ $p_{H2} \ll 0.0001$	0.2176 $p_{H1} = 0.92$ $p_{H2} = 0.82$
Genes(4)	*TUBA4A*	*NCBP2* *POLR2D*	*RPS15A* *SMARC4A* *TUBA4A*	*ACTL6A* ***HSP90AB1*** *SMARC4A* *SMARCB1*	*ACTL6A* *IFITM1* *POLR2D* *SMARC4A* ***SYNCRIP***	*ACTL6A* ***HSP90AB1*** *IFITM1* *POLR2D* *SMARC4A* *RPS2*	Supplemental Data S4, subnetwork 4

Significant Targets in the Developmental Process Subnetwork Are Coordinately Regulated

We used a label-free mass spectrometry approach (see "Experimental Procedures") to verify the relative expression of four of the significant targets in the developmental process subnetwork (Fig. 7, *panel 1*) in a new cohort of clinical tissue samples. The differential expression of *IGFBP3* was determined by Western blot. The relative expression change between normal and stage D at the level of mRNA for each of these targets was computed from the microarray and used for comparison. Most all of the targets were up-regulated in cancer in all patients at the level of mRNA and protein (Fig. 8). Overall these data indicate coordinated regulation at the level of mRNA and protein but also highlight the relatively large variation of expression of both mRNA and protein across patients. An interpretation consistent with this observation is that subtle changes in the transcription of one or more targets may have a synergistic effect on the activity of the other targets in maintaining the phenotype, something the measure of mutual information is well suited to capture and that is consistent with our guiding hypothesis.

The Advantages of an Integrated -Omics Approach to Cancer

Colon cancer has a strong genetic basis due to the accumulation of somatic mutations in oncogenes and tumor

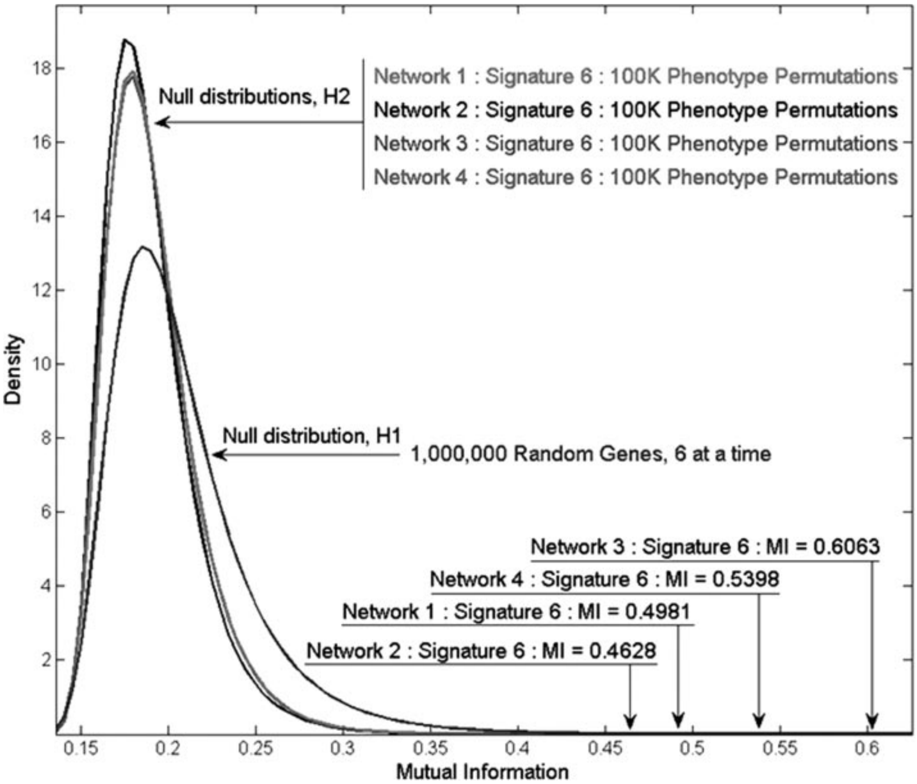


FIG. 6. **Estimated null distributions (probability density function) for hypotheses H1 and H2.** For the H1 null distribution, an array of pseudo, six-gene signatures was computed from 1,000,000 combinations of genes randomly selected six at a time from the decoy data set. The activity value for each pseudosignature was computed between normal and stage D followed by the MI scores, which were then used to populate the distribution. For the H2 null distributions, as each signature (best six) comprises different genes, a separate H2 null distribution was computed for each. First we computed an array of 100,000 (*100K*) random permutations of phenotypes ($n = 171$). Then using the six genes for each signature (Table III), we computed an activity value for 16 pseudonormals and 50 pseudostage D samples (refer to array in Fig. 5). We then computed 100,000 MI scores to populate the H2 distributions. H1 and H2 were modeled by the generalized extreme value distribution in Matlab.

suppressor genes. However, it is also widely accepted that because of the resiliency of mammalian cells single gene mutations are usually insufficient to cause this disease (34). Although a great deal of work has been done identifying genes involved in colon cancer as well as the canonical pathways to which they resolve (35–37), comparatively little work has been done to evaluate the functional protein interactions derived directly from proteomics data. It is in fact not known how genomics, transcriptomics, or proteomics perspectives may differently inform our understanding of colon cancer onset or progression. Classification of disease phenotype using candidate gene, candidate RNA, or candidate protein target approaches has been the bedrock of modern -omics research. However, in some cases these single gene/protein models of disease have been disappointing in follow-on studies (3). Alternative approaches, using network and subnetwork classifiers, are currently under examination. In this study, we searched for protein subnetworks by leveraging a database built on a very large number of legacy experiments using proteomics data as a seed to discover subnetworks discriminative of late stage CRC. It was thought that this approach would quickly lead us to significant protein-protein interaction subnetworks that would reveal the functional cause, or consequence, of stage-specific phenotype(s). We then developed a novel approach for searching within these subnetworks for particular signatures that are significant discriminators of the disease phenotype using a scoring process based on gene expression data. Computational and biostatistical methods to unify proteomics and gene expression data are a powerful way of identifying novel interaction signatures involved in late stage cancer. Although gene lists combined with rigorous statistical analysis can identify significant genes whose expression profiles cluster, the resultant gene lists alone provide no functional information of the post-transcriptional mechanism(s) of dysregulation.

One criticism of our approach is that gel-based proteomics experiments, which provided the seed proteins for our search, typically identify highly abundant proteins as differentially expressed. Many of these are either so-called “housekeeping” genes with a role in metabolism or in any case may often be considered unimportant to a disease phenotype such as cancer as they may lack transcription factors or receptors as protein classes. However, our integrated approach was able to locate these seed proteins within regulatory subnetworks of great interest. Each of the four subnetworks we scored included between eight and 12 proteomic targets that were directly identified. This underscores the usefulness of a network-based approach that identifies specific and significant functional interactions possibly relevant to the pathophysiology of cancer. It also revealed the large diversity of subnetwork interactions in which these high expressers are evidently involved. However, as the scoring shows, the entire subnetwork(s) is not statistically significant for classifying phenotype. Only the end product of our quantitative approach, which identified

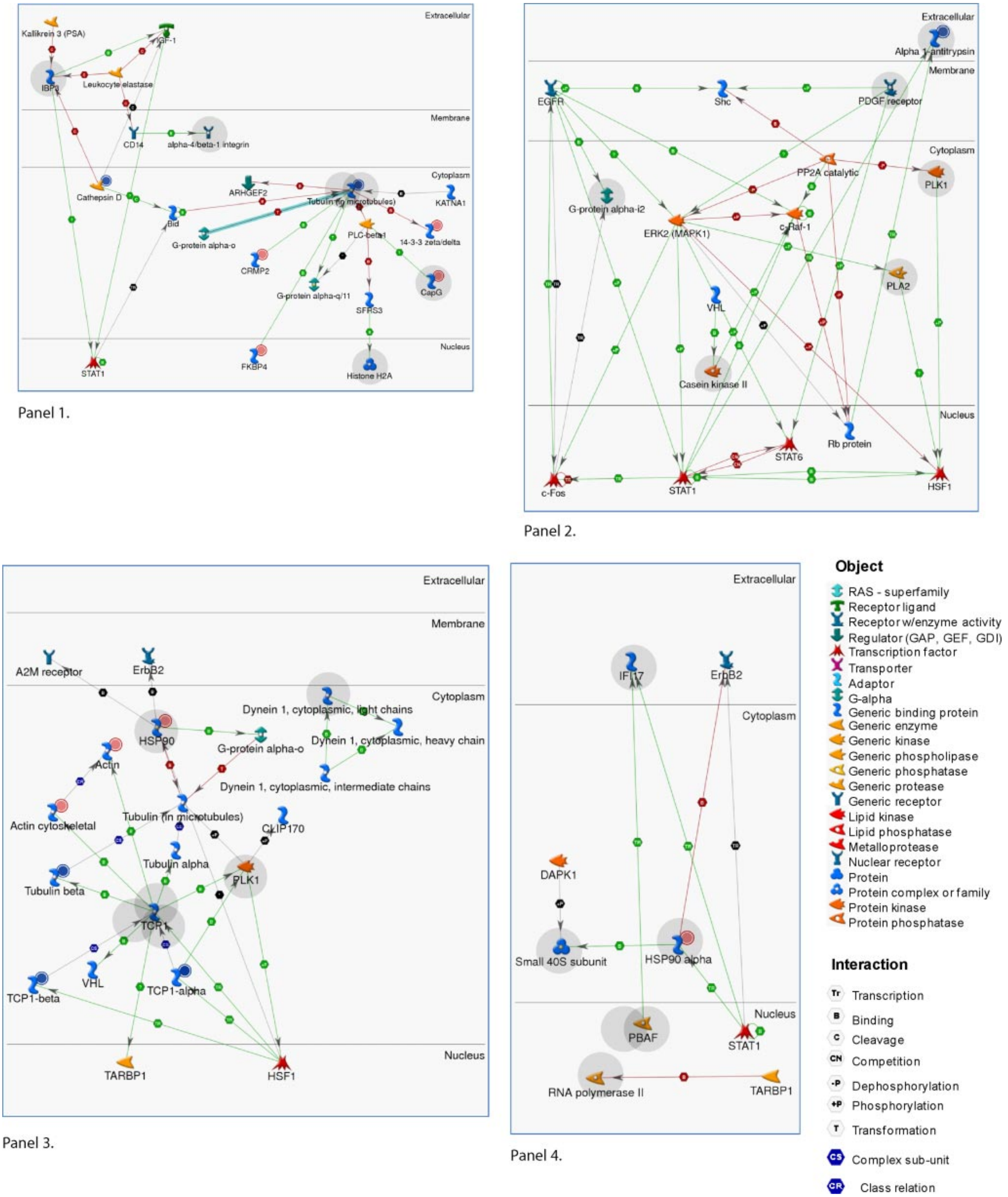


FIG. 7. **Expanded subnetworks from corresponding signatures.** Signature 6 proteins were expanded by one hop inside the corresponding parent subnetwork(s) (Fig. 4 and supplemental Data S4) to infer functional relevance. Signature 6 proteins are highlighted *gray*; *overlapping gray circles* indicate multiple members or subunits of a complex participating in the signature. See also Table III, column labeled Signature 6. *Horizontal lines* demark cellular compartments. *Panels 1–4* are pruned versions of the parent subnetworks, Fig. 4 and supplemental S4 (subnetworks 2–4), respectively.

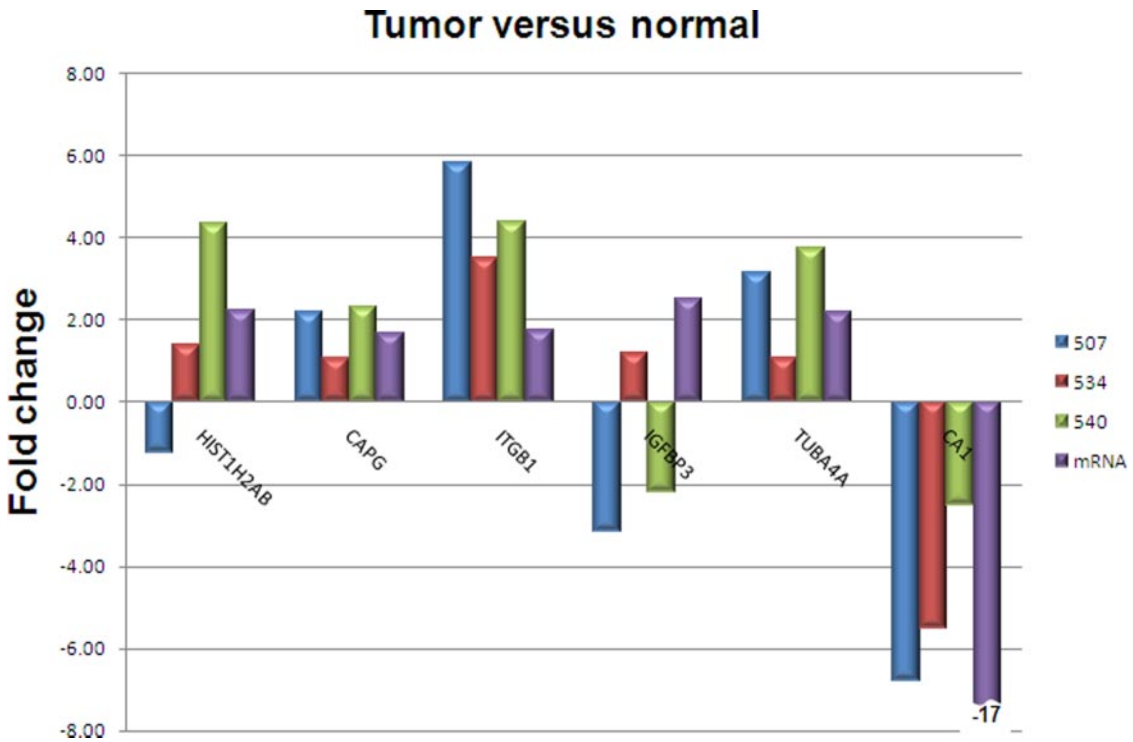


FIG. 8. **Relative expression change of signature proteins and mRNA in subnetwork 1, tumor *versus* normal, for three patients (507, 534, and 540).** mRNA values are the difference between the means measured using the normal samples (n = 16) and stage D samples (n = 50) obtained from the microarray. Protein -fold change was determined by label-free mass spectrometry except for IGFBP3, which was determined by Western blot. As most targets were up-regulated in the tumor, carbonic anhydrase I (*CA1*) is included as a loading control to indicate that the observed up-regulation of most targets was not merely due to a greater amount of tumor digest on column *versus* normal.

root nodes with functional interactions significant for phenotype, presents a focused set of testable hypotheses suitable for validation by perturbation experiments. Additionally we acknowledge that different protein interaction databases are likely to return different subnetworks given the same target seed. However, this is most likely because present day databases represent an undersampling of the human interactome, coverage of which has been recently estimated at less than 1% (38), and not because of any inherent arbitrariness attributable to our approach. Indeed as coverage of the human interactome continues to improve, interaction databases are likely to converge with respect to subnetwork selectivity.

Using the measure of mutual information to score the networks had an advantage over other classification methods in that there is no requirement that the underlying data be normally distributed. This made the method particularly well suited to examining gene expression data that, for many of the genes in our networks, exhibited non-normal distributions of expression for particular stages of cancer. Pairing this approach with exhaustive combinatorial search, *versus* a greedy search, reduces the possibility that the signatures represent a local rather than global maximum. A complete exhaustive search of the expression landscape for even larger combinations of gene products (>6) is limited only by computer power. Finally some of the proteins identified in our signatures did not, independently, have a significant change in mRNA expression at least not enough to be considered significant by simple gene expression profiling. But it is certainly conceivable that the cumulative effect of small changes in network activity (mRNA expression) may lead to significant changes in the proteome, and this was a guiding hypothesis of our study.

As high throughput methods continue to produce more genomics and proteomics data, it will become increasingly important to find new ways to integrate these data and to provide precise, quantitatively significant classifications of human disease stage. These classifiers will likely be critical to the assessment of individual phenotype important for development of personalized medicine.

* This work was supported, in whole or in part, by National Institutes of Health Grants UL1-RR024989 from the National Center for Research Resources (Clinical and Translational Science Awards), P30-CA043703, CWRU Cancer Research Center Proteomics core, and T32-GM008803 from the NIGMS (Institutional National Research Service Award).

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

¶ To whom correspondence should be addressed: Dept. of Pharmacology, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106. Tel.: 216-368-4014; Fax: 216-368-6846; E-mail: rkn6@case.edu.

REFERENCES

1. Vidal, M. (2005) Interactome modeling. *FEBS Lett.* **579,** 1834–1838
2. Edelman, E. J., Guinney, J., Chi, J. T., Febbo, P. G., and Mukherjee, S. (2008) Modeling cancer progression via pathway dependencies. *PLoS Comput. Biol.* **4,** e28
3. Auffray, C. (2007) Protein subnetwork markers improve prediction of cancer outcome. *Mol. Syst. Biol.* **3,** 141–142
4. Shoemaker, B. A., and Panchenko, A. R. (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.* **3,** e42
5. Ries, L. A. G., Melbert, D., Krapcho, M., Mariotto, A., Miller, B. A., Feuer, E. J., Clegg, L., Horner, M. J., Howlader, N., Eisner, M. P., Reichman, M., and Edwards, B. K. (eds) (2007) *SEER Cancer Statistics Review, 1975–2004*, National Cancer Institute, National Institutes of Health Bethesda, MD
6. Friedman, D. B., Hill, S., Keller, J. W., Merchant, N. B., Levy, S. E., Coffey, R. J., and Caprioli, R. M. (2004) Proteome analysis of human colon cancer by two-dimensional difference electrophoresis and mass spectrometry. *Proteomics* **4,** 793–811
7. Bi, X., Lin, Q., Foo, T. W., Joshi, S., You, T., Shen, H. M., Ong, C. N., Cheah, P. Y., Eu, K. W., and Hew, C. L. (2006) Proteomic analysis of colorectal cancer reveals alterations in metabolic pathways. *Mol. Cell. Proteomics* **6,** 1119–1130
8. Mazzanti, R., Solazzo, M., Fantappié, O., Elfering, S., Pantaleo, P., Bechi, P., Cianchi, F., Ettl, A., and Giulivi, C. (2006) Differential expression proteomics of human colon cancer. *Am. J. Physiol.* **290,** G1329–G1338
9. Alfonso, P., Nunez, A., Madoz-Gurpide, J., Lombardia, L., Sanchez, A., and Casa, L. (2005) Proteomic expression analysis of colorectal cancer by two-dimensional differential gel electrophoresis. *Proteomics* **5,** 2602–2611
10. Rahman-Roblick, R., Roblick, U. J., Hellman, U., Conrotto, P., Liu, T., Becker, S., Hirschberg, D., Jörnvall, H., Auer, G., and Wiman, K. G. (2007) p53 targets identified by protein expression profiling. *Proc. Natl. Acad. Sci. U. S. A.* **13,** 5401–5406
11. Volmer, M. W., Stühler, K., Zapatka, M., Schöneck, A., Klein-Scory, S., Schmiegel, W., Meyer, H. E., and Schwarte-Waldhoff, I. (2005) Differential proteome analysis of conditioned media to detect Smad4 regulated secreted biomarkers in colon cancer. *Proteomics* **5,** 2587–2601
12. Tan, S., Seow, T. K., Liang, R. C., Koh, S., Lee, C. P., Chung, M. C., and Hooi, S. C. (2002) Proteome analysis of butyrate-treated human colon cancer cells (HT-29). *Int. J. Cancer* **98,** 523–531
13. Ahmed, N., Oliva, K., Wang, Y., Quinn, M., and Rice, G. (2003) Proteomic profiling of proteins associated with urokinase plasminogen activator receptor in a colon cancer cell line using an antisense approach. *Proteomics* **3,** 288–298
14. Ekins, S., Nikolsky, Y., Bugrim, A., Kirillov, E., and Nikolskaya, T. (2006) Pathway mapping tools for analysis of high content data. *Methods Mol. Biol.* **356,** 319–350
15. Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3,** 140
16. Viswanathan, S., Unlü, M., and Minden, J. S. (2006) Two-dimensional difference gel electrophoresis. *Nat. Protoc.* **1,** 1351–1358
17. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57,** 289–300
18. Williams, A. C., Smartt, H., H-Zadeh, A. M., Macfarlane, M., Paraskeva, C., and Collard, T. J. (2007) Insulin-like growth factor binding protein 3 (IGFBP-3) potentiates TRAIL-induced apoptosis of human colorectal carcinoma cells through inhibition of NF-κB. *Cell Death Differ.* **14,** 137–145
19. Slattery, M. L., Samowitz, W., Curtin, K., Ma, K. N., Hoffman, M., Caan, B., and Neuhausen, S. (2004) Associations among IRS1, IRS2, IGF1, and IGFBP3 genetic polymorphisms and colorectal cancer. *Cancer Epidemiol. Biomark. Prev.* **13,** 1206–1214
20. Stallmach, A., von Lampe, B., Matthes, H., Bornhöft, G., and Riecken, E. O. (1992) Diminished expression of integrin adhesion molecules on human colonic epithelial cells during the benign to malign tumour transformation. *Gut* **33,** 342–346
21. Andreu, P., Colnot, S., Godard, C., Laurent-Puig, P., Lamarque, D., Kahn, A., Perret, C., and Romagnolo, B. (2006) Identification of the IFITM family as a new molecular marker in human colorectal tumors. *Cancer Res.* **66,** 1949–1955
22. Barker, N., Hurlstone, A., Musisi, H., Miles, A., Bienz, M., and Clevers, H. (2001) The chromatin remodeling factor Brg-1 interacts with β-catenin to promote target gene activation. *EMBO J.* **20,** 4935–4943
23. Kitadai, Y., Sasaki, T., Kuwai, T., Nakamura, T., Bucana, C. D., Hamilton, S. R., and Fidler, I. J. (2006) Expression of activated platelet-derived growth factor receptor in stromal cells of human colon carcinomas is associated with metastatic potential. *Int. J. Cancer* **119,** 2567–2574
24. Izeradjene, K., Douglas, L., Delaney, A., and Houghton, J. A. (2005) Casein kinase II (CK2) enhances death-inducing signaling complex (DISC) activity in TRAIL-induced apoptosis in human colon carcinoma cell lines. *Oncogene* **24,** 2050–2058
25. Tapia, J. C., Torres, V. A., Rodriguez, D. A., Leyton, L., and Quest, A. F. (2006) Casein kinase 2 (CK2) increases survivin expression via enhanced β-catenin-T cell factor/lymphoid enhancer binding factor dependent transcription. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 15079–15084
26. Mounier, C. M., Wendum, D., Greenspan, E., Fléjou, J. F., Rosenberg, D. W., and Lambeau, G. (2008) Distinct expression pattern of the full set of secreted phospholipases A2 in human colorectal adenocarcinomas: sPLA2-III as a biomarker candidate. *Br. J. Cancer* **98,** 587–595
27. Watari, A., Takaki, K., Higashiyama, S., Li, Y., Satomi, Y., Takao, T., Tanemura, A., Yamaguchi, Y., Katayama, I., Shimakage, M., Miyashiro, I., Takami, K., Kodama, K., and Yutsudo, M. (2006) Suppression of tumorigenicity, but not anchorage independence, of human cancer cells by new candidate tumor suppressor gene CapG. *Oncogene* **25,** 7373–7380
28. Holtrich, U., Wolf, G., Bräuninger, A., Karn, T., Böhme, B., Rübsamen-Waigmann, H., and Strebhardt, K. (1994) Induction and down-regulation of PLK, a human serine/threonine kinase expressed in proliferating cells and tumors. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 1736–1740
29. Takahashi, T., Sano, B., Nagata, T., Kato, H., Sugiyama, Y., Kunieda, K., Kimura, M., Okano, Y., and Saji, S. (2003) Polo-like kinase 1 (PLK1) is overexpressed in primary colorectal cancers. *Cancer Sci.* **94,** 148–152
30. Weichert, W., Kristiansen, G., Schmidt, M., Gekeler, V., Noske, A., Niesporek, S., Dietel, M., and Denkert, C. (2005) Polo-like kinase 1 expression is a prognostic factor in human colon cancer. *World J. Gastroenterol.* **11,** 5644–5650
31. Schmidt, M., Hofmann, H. P., Sanders, K., Sczakiel, G., Beckers, T. L., and Gekeler, V. (2006) Molecular alterations after Polo-like kinase 1 mRNA suppression versus pharmacologic inhibition in cancer cells. *Mol. Cancer Ther.* **5,** 809–817
32. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y. E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M. H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., and Stratton, M. R. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158
33. Chen, Y., Choong, L. Y., Lin, Q., Philp, R., Wong, C. H., Ang, B. K., Tan, Y. L., Loh, M. C., Hew, C. L., Shah, N., Druker, B. J., Chong, P. K., and Lim, Y. P. (2007) Differential expression of novel tyrosine kinase substrates during breast cancer development. *Mol. Cell. Proteomics* **6,** 2072–2087
34. Vogelstein, B., and Kinzler, K. W. (2004) Cancer genes and the pathways they control. *Nat. Med.* **10,** 789–799
35. Zou, T. T., Selaru, F. M., Xu, Y., Shustova, V., Yin, J., Mori, Y., Shibata, D., Sato, F., Wang, S., Olaru, A., Deacu, E., Liu, T. C., Abraham, J. M., and Meltzer, S. J. (2002) Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene* **21,** 4855–4862
36. Williams, N. S., Gaynor, R. B., Scoggin, S., Verma, U., Gokaslan, T., Simmang, C., Fleming, J., Tavana, D., Frenkel, E., and Becerra, C. (2003) Identification and validation of genes involved in the pathogenesis of colorectal cancer using cDNA microarrays and RNA interference. *Clin. Cancer Res.* **9,** 931–946

37. Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., and Vogelstein, B. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113

38. Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 6959–6964

39. Marouga, R., David, S., and Hawkins, E. (2005) The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Anal. Bioanal. Chem.* **382,** 669–678